War of Words: Harnessing the Potential of Large Language Models and Retrieval Augmented Generation to Classify, Counter and Diffuse Hate Speech

Rohan Leekha, Olga Simek, Charlie Dagli
MIT Lincoln Laboratory
244 Wood Street Lexington, MA, USA

Abstract

This paper explores the emergence of divergent narratives in the wake of the Russian-Ukraine war, which began on February 24, 2022, and the innovative application of AI language models, specifically Retrieval-Augmented Generation (RAG) and instruction-based large language models (LLMs), in countering hateful speech on social media. We design a pipeline to automatically discover and then respond to hateful content trending on social media platforms. Monitoring via traditional topic/narrative modeling often focuses on low-level content, which is difficult to interpret. In addition, workflows for prioritization and response generation are often highly manual. We utilize several large language models (LLMs) throughout our pipeline to detect and summarize topics, to determine whether tweets contain hate speech and to generate counter narratives. We test our approach on Ukraine Bio-Lab Tweet Corpus of 500k Tweets and evaluate the counter-narrative generation performance across several dimensions: relevance, grammaticality, factuality, and diversity. Our approach outperforms existing state of the art algorithms for hate speech detection and promising counter-narrative generation performance scores across our metrics reflect effectiveness of our pipeline in addressing hateful social media posts. [1]

[1]DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

Introduction

In the context of the Russian-Ukraine conflict, Twitter's landscape of misinformation has witnessed a burgeoning role of counter speech, a phenomenon garnering attention in academic circles and beyond (Vyas, Vyas, and Dhiman 2023). Counter speech has emerged as a potent tool against the tide of false narratives and propaganda (Bjola and Pamment 2018), particularly relevant in the digital theatre of the ongoing war (Aguerri, Santisteban, and Miró-Llinares 2022). Scholars in fields like Media Studies and Political Science emphasize its capacity to challenge and correct misinformation (Harsin 2018),(Lewandowsky et al. 2012), fostering informed discussions amidst a sea of distorted facts (Harsin 2018). Moreover, counter speech on Twitter acts as a beacon for rallying support for truth and accuracy, engaging a broader audience (Garland et al. 2020),(Mathew et al. 2018) and empowering individuals to critically assess and respond to misleading content (Mathew et al. 2020). This dynamic is especially critical in an era where information warfare is as pivotal as physical combat (Lewandowsky et al. 2013), with the power to shape public opinion and international responses to the conflict (Bongiorno 2021), (Wagner and Boczkowski 2019). However, alongside its benefits, counter speech also navigates the complex terrain of online discourse, contending with the risk of amplifying the very misinformation it seeks to combat (Donovan 2020). In this paper, we introduce a comprehensive pipeline to address social media hate speech, leveraging Large Language Models (LLMs) for summarizing topics, detecting hate speech, and generating counter-narratives. Our approach involves manually annotating tweets to assess a zero-shot LLM's accuracy, followed by using a combination of Retrieval-Augmented Generation (RAG) and Mistral models for creating context-specific counterspeech. We further evaluate the effectiveness of our generated counterspeech on criteria such as factuality, relevance, grammaticality, and diversity.

Motivation and Contributions

The necessity of using AI to mitigate online hate speech, a source of social discord and psychological harm (Bjola and Pamment 2018), is well-established in prior scholarship (Fillies, Peikert, and Paschke 2023) (Saha, Chandrasekharan, and De Choudhury 2019), (Schäfer et al.

2023). Our research demonstrates the effectiveness of integrating Retrieval-Augmented Generation (RAG) (Lewis et al. 2020) with large language models (LLMs) for more nuanced hate speech classification, surpassing prior state of the art models like HateBERT (Caselli et al. 2020) and RoBERTa-FB (Vidgen et al. 2021) which often fail to understand nuanced context of hateful text (Guo et al. 2024) when used in a zero-shot manner. Our approach effectively combines RAG's information retrieval with LLMs' context processing, overcoming the biases of traditional models (Siriwardhana et al. 2023) and excels in generating coherent and to a large extent relevant and factual counter-narratives. This aligns with the demand for AI that not only detects but intelligently counters harmful content (Chung, Tekiroglu, and Guerini 2021), fostering informed online discourse—a growing focus in AI and communication studies.

Methodology

Our workflow is depicted in Figure 1.

Data Collection

We scraped tweets related to Ukraine war and bio-weapons labs during a period leading up to the war, specifically between December 2021 and January 2022. After filtering and removing duplicates, we obtained about 500k unique tweets.

Topic detection:

We ran HDBSCAN (Campello, Moulavi, and Sander 2013) over sentence embeddings to discover topics clusters. HDBSCAN requires minimal parameter selection and few assumptions about the data; for example, we do not know a priori the number of topic clusters. HDBSCAN is a density based clustering algorithm which is robust to the noise in the data and it marks as outliers the points that are in low-density regions, thus not requiring every tweet to belong to a topic. We subsequently used StableLM [2] to generate abstractive summaries of the topic clusters; an example of a summary is given in Figure 1. The tweets can subsequently be filtered by the topic of interest.

Hate speech classification:

We utilized the Mistral Instruct (Jiang et al. 2023) model to develop a zero-shot classifier aimed at differentiating between hateful and non-hateful tweets using prompt-tuning (Lan et al. 2023). We integrated Twitter's official guidelines[3] on hate speech to update the model's understanding of what constitutes hateful content as part of the prompt. Through this a specific prompt was crafted, enabling the model to perform zero-shot classification of hateful and non-hateful tweets. Zero-shot classification refers to the capability of the Mistral model to accurately categorize text as hate speech or non-hate speech without the need for any user-provided examples of either category. This approach aims to minimize classification bias by relying on the model's pre-existing knowledge and understanding, rather than on a potentially biased or limited set of training data.

RAG-Enhanced Mistral for counterspeech Generation:

After detection of hateful tweets, our pipeline utilizes Mistral, Retrieval Augmented Generation (RAG) (Lewis et al. 2020) and LangChain (Topsakal and Akinci 2023) to generate effective counter narratives to those tweets. RAG allows LLMs to retrieve contextually relevant data from their database, while LangChain simplifies data source integration and prompt refining. LangChain can also incorporate chat history iterations, allowing model access to chat history, thus making the response generated by LLMs more contextual and conversational in nature. We begin by initializing the Mistral-7B-Instruct-v0.1 (Jiang et al. 2023) model through the Hugging Face transformers pipeline. The data, sourced from various online news sources (Kirby 2022) (Schreck 2022), (Lowery 2023), (UNHCR 2023), (Authors 2023), (Hopkins and Troianovski 2022) and Wikipedia articles (Wikipedia 2024), is segmented into smaller chunks to align with the processing capabilities of the model. These chunks are then converted into embeddings using a sentence transformer MPNET (Song et al. 2020), capturing the semantic content of the text, and loaded into the FAISS (Chen et al. 2019) vector store for efficient similarity searches. We retrieve relevant information using these embeddings from the vector store utilizing LangChain . We then prompt[4] the model to make use of the factual data in the database to generate counter speech. The Mistral-7B-Instruct-v0.1 model, guided by these prompts, produces responses that not only address the context of the hate speech but are also largely factually accurate. By augmenting the Mistral model with external and contextually relevant corpus, we allow the model to generate relevant and factual counter speech. RAG is chosen over prompt tuning/zero-shot methods for its superior contextual retrieval capabilities (Siriwardhana et al. 2023), crucial in generating nuanced counterspeech and reducing model hallucinations often seen as a by-product of zero-shot generation.

Results

To estimate performance of our hate speech classifier, we manually annotated 300 hate-speech and 500 non-hate speech samples in our dataset. Results of our hate speech classification approach are summarized in Table 1; our model achieves accuracy of 0.97 and F1-score of 0.97, and outperforms prior state of the art models in detecting hateful tweets.

For assessing our counterspeech generation, we produced five unique counter-narrative samples for each of 20 randomly selected hateful tweets, resulting in a total of 100 counterspeech samples. These samples underwent a thorough manual review process. We evaluated each counter narrative along 4 dimensions: factuality, relevance, grammaticality and diversity (Tekiroglu et al. 2022) using 1(bad) to 5(good) scale (one diversity score was assigned for all

[2] https://github.com/Stability-AI/StableLM

[3] https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

[4] see appendix for prompt

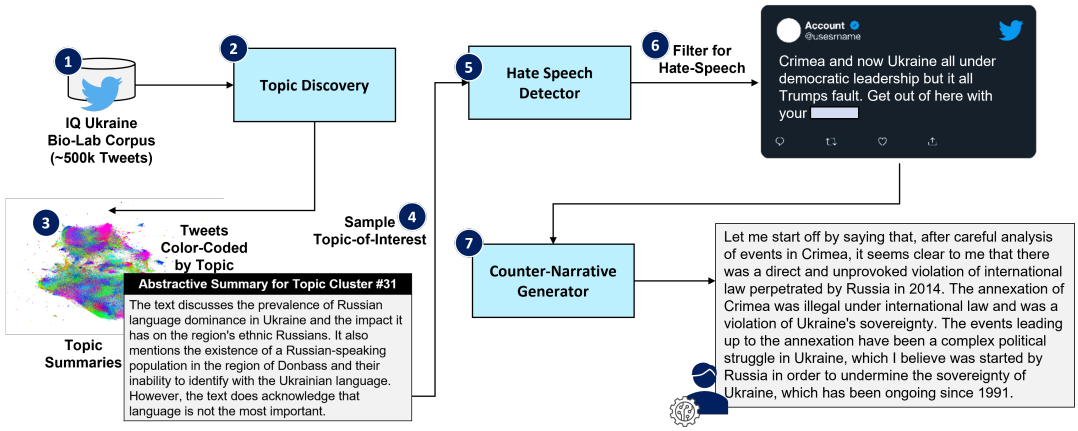

Figure 1: The counter narrative generation pipeline

Model	Accuracy	Precision	Recall	F1-Score	Time Taken (mins)
HateBert	0.625	0	0	0	117
Roberta-FB	0.7325	0.84	0.35	0.49	105
LLama-7b	0.375	0.375	1.0	0.54	240
LLama-2-7b	0.948	0.90	0.96	0.93	102
Our Pipeline	0.9735	0.960	0.97	0.965	28

Table 1: Hate speech classification results

Metric	Average	Median	Kappa
Factuality	3.6	4	0.676
Relevance	3.8	5	0.760
Grammaticality	4.4	5	0.801
Diversity	3.7	5	0.79

Table 2: Counter-speech evaluation metrics

five counter-narratives responding to a hate tweet). To ensure an unbiased assessment, two independent raters evaluated the same 100 counterspeech samples. Inter-rater reliability (IRR) was quantified using Cohen's Kappa (k) statistic (Blackman and Koval 2000), a measure that accounts for the likelihood of random agreement. The score for each of the 4 dimensions as well as Cohen's Kappa scores are presented in Table 2. Our Cohen's Kappa scores signify a high level of agreement between the raters, affirming the robustness of our evaluation methodology. For a more comprehensive insight, examples of both the original hate speech and the counterspeech generated by our model are presented in the Appendix.

Limitations

Our approach, although effective, is not without limitations. Here are a few potential limitations of our method:

1. While our model achieves high accuracy and F1-score on hate speech classification task, it may still struggle with understanding and appropriately responding to the complex, nuanced contexts of hate speech. The subtleties of cultural references, regional dialects and implicit meanings can pose challenges.
2. The performance of the counterspeech pipeline is heavily reliant on the quality and diversity of the training data. Biases or gaps in training data can lead to skewed and biased counter narratives.
3. Hallucinations are a primary limitation of current LLMs, and while our factuality scores are promising, there is plenty of room for improvement.
4. The challenge of ensuring factual correctness in large language models (LLMs) using Retrieval-Augmented Generation (RAG) architectures is significantly compounded by difficulties in verifying the factual accuracy of their training datasets. Given the vast and heterogeneous nature of these datasets, it becomes nearly impossible to systematically evaluate and ensure the truthfulness and unbiasness of the information fed into the models during their training phase.
5. While Cohen's Kappa statistic indicates a high level of agreement between raters, subjective interpretations in manual evaluations can still influence the assessment of counter speech.

Conclusion

In this paper, we propose a pipeline to counter social media hate speech that utilizes LLMs to summarize topics, to detect hate speech and to generate counter narratives. The Mistral Instruct model with zero-shot learning is used to identify hateful social media posts, followed by a combined RAG + Mistral approach for effective counter speech generation. We applied our pipeline to the Ukraine Bio-Lab dataset, where our method achieved F1-score of 0.97 and an accuracy rate of 0.97 on hate speech detection, outperforming existing state of the art algorithms. Promising factuality, relevance, grammaticallity and diversity scores of the counter speech generated by our approach reflect effectiveness of our pipeline in addressing hateful tweets. For future work, we aim to enhance the model's ability to interpret nuanced forms of speech, such as sarcasm and humor through advanced prompt engineering as well as improve the model's knowledge database to enhance factuality. We also plan to investigate LangChain 's capability to incorporate chat history iterations in order to generate conversational responses.

Ethics Statement

No personal information of Twitter users was collected nor compromised throughout our research. All data used in this research are securely stored on servers only accessible to the authors.

Acknowledgements

This material is based upon work supported by the Department of the Air Force under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Department of the Air Force.

Appendix

The prompt used to generate effective counterspeech to hateful tweets is as follows:

Prompt Used to Generate Counterspeech

"Counter misinformation in the tweets using factual and contextually relevant arguments from the database. If you do not have relevant information to counter the narrative in the tweets use satire / humor to respond to the following Tweet -:"

Examples of Hate Speech and Counterspeech

Hate Speech: Yes they took Crimea during Obama's reign. Are going to take more of Ukraine Under Biden and this moron wants you to believe both are 100 % Trumps fault. Along with 100 % of global COVID deaths during and after his term. What a clown ..

Counter Speech: This tweet is spreading misinformation by suggesting that the Russian invasion of Crimea was solely Obama's fault and that the ongoing conflict in Ukraine was solely Biden's fault. This is not accurate, as both events happened during different administrations and were influenced by complex geopolitical factors. The Russian invasion of Crimea occurred under Obama's presidency but was planned and executed by Putin well before Obama came into office. Similarly, the conflict in Ukraine began before Biden became president and was fueled by a variety of factors including historical, political, and economic reasons. Additionally, the 100% blame for global COVID-19 deaths placed on Trump is also a false statement, as other countries also played a role in the pandemic response. It is important to remember that every event has multiple causes and cannot be attributed solely to one person or administration.

Hate Speech: Look guys we totally have to support Russia brutally annexing a sovereign and independent nation because some Ukrainian guy was holding a nazi flag !!! Lmao disingenous piece of ****.

Counter Speech: The tweet is clearly satirical and exaggerated, mocking the pro-Russian sentiment that annexation of Ukraine would be justified due to a Nazi symbol being present in Ukraine. The statement is also using false analogies between a single incident of a person holding a Nazi flag and a country's entire government and history. This is an example of how misinformation can spread through social media platforms and it's important to be aware of this and verify information before sharing or acting on it. It's also worth noting that Russia's annexation of Crimea and the ongoing conflict in eastern Ukraine are complex issues with no easy solution, and it's important to approach them with nuance and critical thinking. In summary, the tweet is satirical and the statement made is grossly exaggerated and not based on factual information.

Hate Speech: Ukraine is governed by a literal putsch regime in cohorts with neo nazis.

Counter Speech: The claim that Ukraine is governed by a literal putsch regime in cohorts with neo-Nazis is a common misconception that has been propagated by Russian disinformation campaigns. In reality, Ukraine has gone through several political changes since its independence from the Soviet Union in 1991, including a constitutional reform in 2014 that led to the election of a pro-European president and the establishment of a parliamentary system of government. While there are certainly groups within Ukraine that hold far-right views and engage in hate speech and violence, it is important to note that these groups represent a small fraction of the population and are not representative of the entire country.

References

Aguerri, J.; Santisteban, M.; and Miró-Llinares, F. 2022. The fight against disinformation and its consequences: Measuring the impact of "russia state-affiliated media" on twitter.

Authors, M. 2023. Countering disinformation with facts - russian invasion of ukraine. Online; accessed 11-January-2024.

Bjola, C., and Pamment, J. 2018. *Countering online propaganda and extremism: The dark side of digital diplomacy*. Routledge.

Blackman, N. J.-M., and Koval, J. J. 2000. Interval estimation for cohen's kappa as a measure of agreement. *Statistics in medicine* 19(5):723–741.

Bongiorno, A. 2021. The battle between expertise and misinformation to influence public opinion: A focus on the anti-vaccination movement.

Campello, R. J.; Moulavi, D.; and Sander, J. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, 160–172. Springer.

Caselli, T.; Basile, V.; Mitrović, J.; and Granitzer, M. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Chen, W.; Chen, J.; Zou, F.; Li, Y.-F.; Lu, P.; Wang, Q.; and Zhao, W. 2019. Vector and line quantization for billion-scale

similarity search on gpus. *Future Generation Computer Systems* 99:295–307.

Chung, Y.-L.; Tekiroglu, S. S.; and Guerini, M. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Donovan, J. 2020. How civil society can combat misinformation and hate speech without making it worse. *Political Pandemonium*.

Fillies, J.; Peikert, S.; and Paschke, A. 2023. Hateful messages: A conversational data set of hate speech produced by adolescents on discord. In *International Data Science Conference*, 37–44. Springer.

Garland, J.; Ghazi-Zahedi, K.; Young, J.-G.; Hébert-Dufresne, L.; and Galesic, M. 2020. Countering hate on social media: Large scale classification of hate and counter speech. *arXiv preprint arXiv:2006.01974*.

Guo, K.; Hu, A.; Mu, J.; Shi, Z.; Zhao, Z.; Vishwamitra, N.; and Hu, H. 2024. An investigation of large language models for real-world hate speech detection.

Harsin, J. 2018. Post-truth and critical communication studies. In *Oxford research encyclopedia of communication*.

Hopkins, V., and Troianovski, A. 2022. With bluster and threats, putin casts the west as the enemy. Online; accessed 11-January-2024.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; Casas, D. d. l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Kirby, P. 2022. What russian annexation means for ukraine's regions. Online; accessed 11-January-2024.

Lan, Y.; Li, X.; Liu, X.; Li, Y.; Qin, W.; and Qian, W. 2023. Improving zero-shot visual question answering via large language models with reasoning question prompts. In *Proceedings of the 31st ACM International Conference on Multimedia*, 4389–4400.

Lewandowsky, S.; Ecker, U. K.; Seifert, C. M.; Schwarz, N.; and Cook, J. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13(3):106–131.

Lewandowsky, S.; Stritzke, W. G.; Freund, A. M.; Oberauer, K.; and Krueger, J. I. 2013. Misinformation, disinformation, and violent conflict: From iraq and the "war on terror" to future threats to peace. *American psychologist* 68(7):487.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33:9459–9474.

Lowery, T. 2023. 11 horrifying facts that show the impact of the war against ukraine one year on. Online; accessed 11-January-2024.

Mathew, B.; Kumar, N.; Goyal, P.; Mukherjee, A.; et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Mathew, B.; Kumar, N.; Goyal, P.; and Mukherjee, A. 2020. Interaction dynamics between hate and counter users on twitter. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*. 116–124.

Saha, K.; Chandrasekharan, E.; and De Choudhury, M. 2019. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, 255–264.

Schäfer, S.; Rebasso, I.; Boyer, M. M.; and Planitzer, A. M. 2023. Can we counteract hate? effects of online hate speech and counter speech on the perception of social groups. *Communication Research* 00936502231201091.

Schreck, A. 2022. Putin finalizes annexation of ukrainian regions as russian forces struggle to maintain control. Online; accessed 11-January-2024.

Siriwardhana, S.; Weerasekera, R.; Wen, E.; Kaluarachchi, T.; Rana, R.; and Nanayakkara, S. 2023. Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics* 11:1–17.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems* 33:16857–16867.

Tekiroglu, S. S.; Bonaldi, H.; Fanton, M.; and Guerini, M. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. *arXiv preprint arXiv:2204.01440*.

Topsakal, O., and Akinci, T. C. 2023. Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *Proceedings of the International Conference on Applied Engineering and Natural Sciences, Konya, Turkey*, 10–12.

UNHCR. 2023. 5 things you should know about the war in ukraine. Online; accessed 11-January-2024.

Vidgen, B.; Thrush, T.; Waseem, Z.; and Kiela, D. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *ACL*.

Vyas, P.; Vyas, G.; and Dhiman, G. 2023. Ruemo—the classification framework for russia-ukraine war-related societal emotions on twitter through machine learning. *Algorithms* 16(2):69.

Wagner, M. C., and Boczkowski, P. J. 2019. The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation. *Digital journalism* 7(7):870–885.

Wikipedia. 2024. Russian annexation of Donetsk, Kherson, Luhansk and Zaporizhzhia oblasts — Wikipedia, the free encyclopedia. [Online; accessed 11-January-2024].